International Journal of Computer Networking,
Wireless and Mobile Communications (IJCNWMC)
ISSN (P): 2250–1568; ISSN (E): 2278–9448
Vol. 12, Issue 1, Jun 2022, 11–20

TRANS STELLAR
•Journal Publications •Research Consultancy

DECISION TREE AND K-NEAREST NEIGHBOR (KNN) ALGORITHMS IN BREAST CANCER DIAGNOSIS: A RELATIVE PERFORMANCE EVALUATION

DHAHI ALSHAMMARI

University of Ha'il, College of Computer Science and Engineering, Department of Computer Science, Kingdom of Saudi Arabia

ABSTRACT

The accelerated growth of databases, which in each of the different fields of knowledge, and among them the health sector, are being created as a response to the evolution, development and articulation of technology in their daily work, The incursion of data mining models have become necessary in response to the need created by this cloud of information. This articulation has made it possible to record a more significant number of attributes related to the same study unit and expand the variety of formats in which data is recorded and stored. This exponential growth of databases has meant that the classic statistical techniques used by experts and researchers cannot fully reveal the underlying information in the set, making it necessary to introduce new analysis techniques such as those initially mentioned. Given this scenario, it is of interest to identify how much additional information the mining models offer on the exploratory analysis of the data, in addition to determining if any mining model will offer additional information. For this, a case study will be carried out on two data sets, each of which seeks to determine the malignancy of a mass detected in the patient's breast based on the characteristics measured on the mass, seeking to support timely decision-making and reducing procedures that can be costly and unnecessary for both the user and the service provider, since experience shows that 70% of the biopsies performed, based on the results of mammography, are unnecessary. Because of the large number of patients, it is critical to rapidly analyse these data to discover disease as early as possible. By altering various parameters of the KNN and decision tree algorithms, breast cancer data from Dr William H. Wolberg's University Hospital of the University of Wisconsin, Madison, Wisconsin-USA with a volume of 700 cases was classified and analysed. Cross-validation was used to compare the training and test data, and the KNN technique had the most effective classification success rate (97.4 per cent) of all the techniques tested.

KEYWORDS: Breast Cancer, Data Mining Techniques, Comparative Analysis, Data Tree,& KNN

Received: Jan 17, 2022; **Accepted**: Feb 07, 2022; **Published**: Mar 02, 2022; **Paper Id.**: IJCNWMCJUN20222

Original Article

1. INTRODUCTION

According to the World Health Organization data, cancer cases led to the death of 9.6 million people in 2018. Breast cancer affects 2.089 million women every year and causes 627 thousand women to die [1]. Breast cancer, which is statistically more common in middle-aged women, is the uncontrolled proliferation of cells in the breast tissue between the milk glands and milk ducts. Unlike benign cancer types, malignant cancer types show the feature of spreading to other tissues. The morphological features of the tumour play an essential role in this differentiation in breast cancer cases. If the mass has a specific structure and the roughness is low, the cancerous tissue is benign, while the margins do not fit into specific criteria. The roughness is high; it is an indicator of the risk of malignant cancer. This distinction is one of the most preferred methods in terms of early diagnosis of the cancer type of patients, improving the treatment processes of the disease and prolonging the life span [2]. The

ability to make this distinction quickly and effectively has led machine learning methods to play an essential role in the medical field [3].

Diagnosis of the type of breast cancer is essential in terms of the course of the disease and the effective implementation of the treatment to be applied. Various methods are used for this purpose. Mammography, needle tip aspiration method and invasive biopsy are the main ones of these methods. In recent years, high accuracy in cancer cases detected by mammography has been achieved by developing image processing algorithms and image acquisition methods. Since the effect of ionising radiation applied to the patient is low and aims to achieve high accuracy, it is preferred in the initial diagnosis stages [4]. The invasive biopsy method is another method of detecting the tumour and its type in the most accurate way. Although implementing this method is more difficult than other methods, it produces more successful results than other methods. In addition, the physiological and psychological effect on the patient causes this method to be used for final diagnosis. The needle-tipped aspiration method, on the other hand, has become the most preferred method as a result of developments in recent years. High accuracy rate and application speed are the most significant advantages of the needle tip aspiration method.

In prediction-based systems, future unknown values are obtained by regression or classification methods, taking the current variables into account. Machine learning is a method that reveals the rules and relations that will enable predictions and define the data with various equations [5]. For this purpose, k-nearest neighbor, decision trees or various hybrid methods are frequently used in the literature [6]. On the other hand, descriptive methods reveal the formations in the data and allow the data to be interpreted more easily. Cluster analyses, association rules and sequential pattern rule frequently use descriptive methods [7].

Among the relevant concepts for developing this work is the exploratory data analysis (EDA, for its acronym in English). The primary purpose of the EDA is to highlight the relevant characteristics of each of the attributes in the data set, using graphical methods, performing classical statistical summaries, identifying distributions in the data, and to study the intensity of the underlying relationship between attributes. The exploratory analysis includes three main phases: the univariate analysis, the bivariate analysis and the multivariate analysis [6]. By exploring the data, significant information can be discovered to solve the problem of interest. This information is obtained from the bounds of the data, descriptive measures of these and their distribution. The information obtained can offer practical advantages in a data mining project, increasing the knowledge and understanding of the attributes involved in the problem [5-7]. Although the EDA is an essential tool in identifying how data is seen, when technology evolves and becomes more accessible to different areas of knowledge, it generates a massive increase in data. The identification of information immersed in them becomes a more complex process in which exploration is not enough[6]. A set of models is then developed, supported by different areas of knowledge, such as mathematics, statistics, and computing, which are complemented by the EDA to reveal the underlying information in these large data clouds.

These models added with the EDA and different validation methods are consolidated to form a data analysis technique known as Data Mining or English Data Mining. It can be said then that data mining is a process used to discover information that remains underlying in a set of historical data [1]. In summary, the term Data Mining refers to the whole process that involves the collection and analysis of data, the development of inductive learning models and the adoption of practical decisions and actions based on the knowledge acquired [6]. In essence, data mining models can be classified into supervised and unsupervised. The decision tree model will be used for the current development, as supervised models and

the K-means cluster as an unsupervised technique.

In this study, the effects of the k-nearest neighbor algorithm's distance criteria and the change of the number of neighborhoods and the split number parameter of the decision tree algorithms on the classification success were compared and shown on the breast cancer dataset. Performance evaluations were made in cases with the highest classification success.

2. MATERIAL AND METHOD

In this study, Breast cancer data collected by William H. at the University of Wisconsin hospital were used. The data were classified by changing the various parameters of the k-nearest neighbor and decision tree algorithms and the performance analyses of the algorithms were compared.

2.1. Dataset Overview

The data set used in the study consists of 10 features obtained by digitising breast cancer images. The dataset containing 700 samples was marked as benign class 2 and malignant class 4 as a diagnostic result. There are 242 (34.55%) malignant and 458 (65.45%) benign samples in the data set [8]. In Table 1, the definitions, value ranges, averages and standard deviation values of 10 features in the data set are given.

Table 1: Attribute Descriptions and Values

Number	Attribute Definition	Value	Average	Standard deviation
1	Closing Thickness	1 to 10	4.4864	2.8482
2	Size Uniformity	1 to 10	3.1815	3.0957
3	Shape isomorphism	1 to 10	3.2472	3.0179
4	Adhesion	1 to 10	2.8684	2.8926
5	Epithelial Dimension	1 to 10	3.2663	2.2452
6	Bare Core	1 to 10	3.5794	3.6794
7	Soft Chromatin	1 to 10	3.4795	3.4835
8	Normal Nucleoli	1 to 10	2.8977	3.0805
9	mitosis	1 to 10	1.6190	1.7493
10	Class	2 to 4		

2.2. K NearestNeighbor

It is a sample-based algorithm used to solve classification problems. It uses the training data to perform the learning process. Assuming k data points are nearby, a distance criterion is used to determine how similar the data are [9]. The Minkowski, Euclid, Chebyshev, and cosine equations can be used to calculate these distances. In the literature, Euclidean distance is often preferred. The distance between $P= x_1,x_2, \dots ,x_n$ and $Q = y_1,y_2, \dots ,y_n$ is calculated as shown in equation 1, where P and Q are two sets of points.

$$\sqrt{\sum_{i=1}^{n}(x_i - y_{i)^2} \dots (Eq.1)}$$

The class labels of the k nearest central neighbours in the trained data set are evaluated to see if they correspond to the new data when it is fed into the classification process. Based on most class labels, a new piece of data is added to the cluster. The simplicity, lack of non-linear training processes, and effectiveness of the K-nearest neighbour approach even when dealing with noisy training data are some of the method's most notable advantages.However, the k-nearest neighbor

algorithm has several disadvantages [11]. The high level of memory usage depending on the size of the data set and the increase in the processing load as the data is thrown make it difficult for the k-nearest neighbor algorithm to work on data with many features. In addition, the sensitivity of the performance measurement to the number of neighbors entered from outside, and the sensitivity to the determined distance criterion are the most basic deficiencies [12].

2.3. Decision Trees

Decision trees are algorithms that create the tree structure starting from the top to classify the data in the first stage. The tree structure is named roots, branches and leaves, provided that it starts from the top. Branches are tied to nodes, provided that each branch is connected to the root above. Each feature in the data represents a node in the tree after classification. The nodes between these tree structures are classification rules and each leaf is considered a class [13].

There are many decision tree algorithms in the literature. The tree formation principles show different classification successes depending on the data diversity. ID3, J48, C4.5 and C5 algorithms are the most well-known decision tree algorithms. The C4.5 algorithm uses the entropy equation to select the most distinguishing feature during the classification process. With the entropy equation, uncertain situations in the data and the rate of randomness in the data can be measured. To representp_1,p_2, …,p_n probability states, the sum of all these states must be 1. Considering this situation, entropy is calculated in equation 2 [14].

$$I(p) = -\sum_{i=1}^{k} p_i \times \log(p_i) \dots (Eq.2)$$

After calculating the entropies of all the features in the database, the Information(D, S) value is calculated as given in equation 3.

$$(\mathrm{D,S})\mathrm{Information} = \sum_{i=1}^{n} \frac{Ti}{T} I \dots (Eq.3)$$

The gain values of the elements whose information values are calculated as given in equation 4. The element with the maximum gain value is placed on the top root node.

$$(D,S)\ Gain\ value = (S)Information - (D,S)Information \dots (Eq.4)$$

Other decision tree algorithms also show high classification success on various data. The J48 algorithm is based on eliminating weak branches to increase the classification success while creating the tree structure [15]. In addition, there are studies in the literature in which hybrid methods to be created with decision trees increase the classification success [16]. Decision tree algorithms are easy to interpret and effective for solving problems with multiple outputs. However, algorithms can produce complex branches and problems such as memorising trees [17].

2.4. Performance Evaluation

In the classification processes made with data mining methods, the success of the classification should be tested with various methods. Some of the data used to train the system can be used as test data, or data with the same attributes as the data set can be used for testing. For this purpose, a certain percentage of the data can be separated as test data or cross-validation methods can be used [18].

The prediction classes produced due to the classification method and the actual classes in the data are expressed in

separate clusters. With these clusters, information about various parameters of the classification result can be obtained. The naming criteria of these clusters are given in Table 2. The intersection of the predicted positive value and the true positive value is true positive (TP), the intersection of the predicted positive value and the true negative value is false positive (FP), the intersection set of the predicted negative value and the true positive value is a false negative (FN). The predicted negative value The intersection set of and true negative value is called true negative (TN). Placing these terminologies in the error matrix is as shown in Table 2.

Table 2: Error Matrix Structure

Actual Value			
		Positive	Negative
Estimated Value	Positive	True Positive	False positive
	Negative	False Negative	True Negative

Some of the performance criteria used to measure the classification success obtained by data mining methods are given in Table 3. The accuracy value indicates how much the values obtained due to the classification process express the actual values. The sensitivity equation reveals the ability of the classification process to detect true positive values. Similarly, the predictive value reveals the classification process's ability against the classification result's negative value. It demonstrates the ability of the classification process with the precision equation to eliminate false positive values. On the other hand, F score value is a calculation method used to eliminate unusual situations that may arise as a result of calculating precision and sensitivity values.

Table 3: Some Performance Measures used In Data Mining

Criterion	Equation
Truth	$\frac{TP + TN}{TP + TN + FP + FN}$
Sensitivity	$\frac{TP}{TP + FN}$
Decisiveness	$\frac{TN}{TN + FP}$
Precision	$\frac{TP}{TP + FP}$
F Score	$2.\frac{Precision * Sensitivity}{Precision \ + Sensitivity}$

3. EXPERIMENTAL RESULTS

A version of open-source Python 3.6 was used for the experiments. K closest neighbour and decision tree methods were used to explore how parameter changes affected classification accuracy.Cross-validation was used for test data. The effect of the distance criterion and the number of neighbors used in the weight value assignment in the k-nearest neighbor algorithm and the maximum separation parameters for the decision tree algorithm on the classification success were compared. In the K-nearest neighbor algorithm, the effect of Euclid, Minkowski, Chebyshev and cosine distance on classification success as a distance criterion and the effect of cluster numbers ranging from 1 to 10 for each distance criterion on classification success are shown. The Gini index was used as the separation criterion in the decision trees, and the classification success was measured by changing the maximum separation number between 1 and 10. The error

matrices of the highest classification successes obtained with the K-nearest neighbor and decision tree algorithm were obtained. The performance evaluations of the most successful results were made due to these matrices.

In Figure 1, the results were obtained from the experiments with the k-nearest neighbor algorithm. In the classification in which all the features in the data set are used, the average results obtained with the cosine distance is higher than the average of the results obtained with the other distance criteria. The most successful classification was obtained in experiments with cosine distance of 4 neighborhoods with 97.30%.

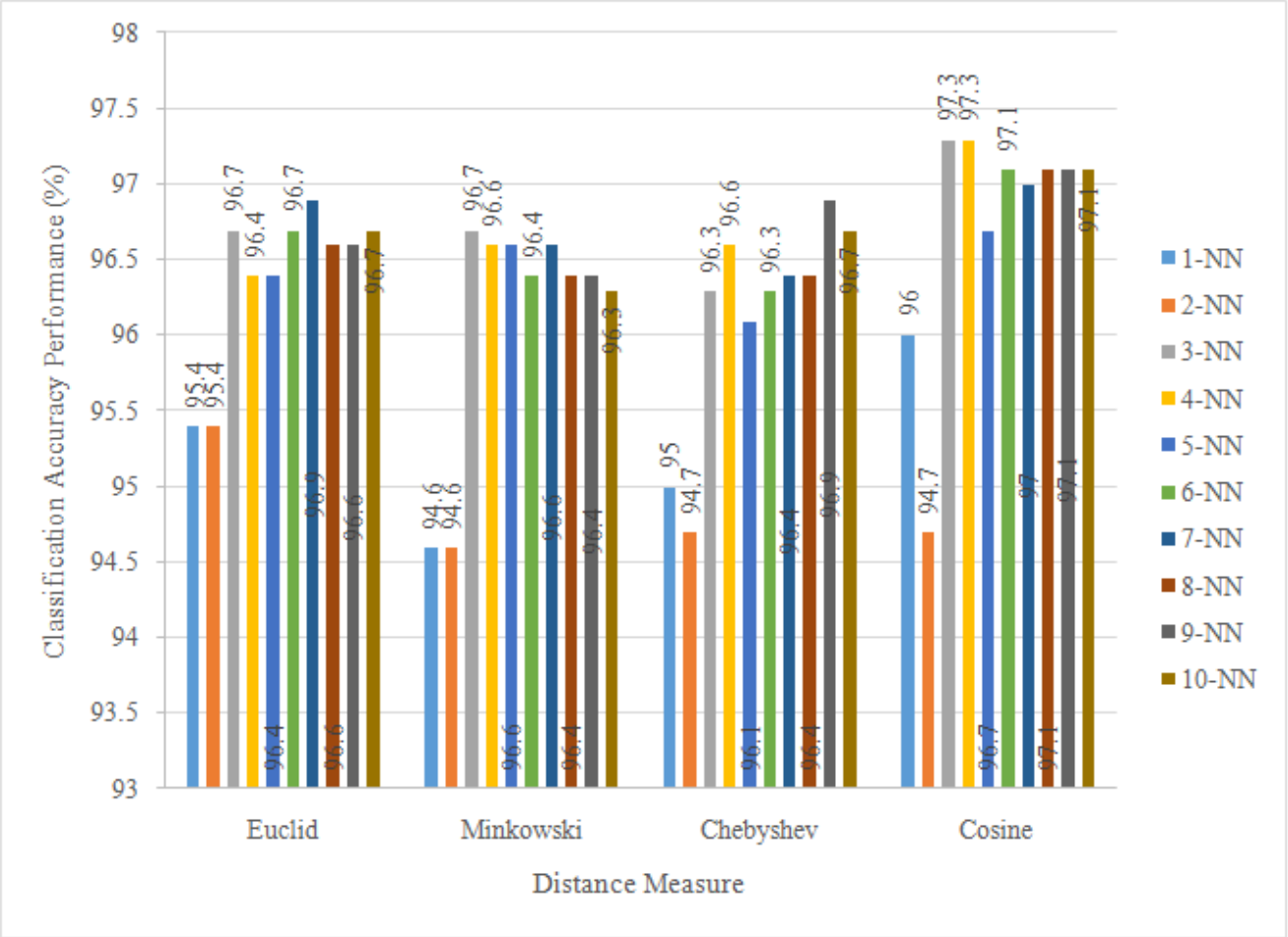

Figure 1: Classification Success According to Distance and Number of Neighbors Parameters

In Table 4, the error matrix of the experiment was most successful with 97.40% in the k-nearest neighbor algorithm.

Table 4: Error Matrix of the Highest Classification Obtained with KNN

		Actual Values	
		2	4
Estimated Values	2	445	6
	4	13	235

In Figure 2, the effect of the number of divisions on the classification success in the decision tree created using the

Gini index.

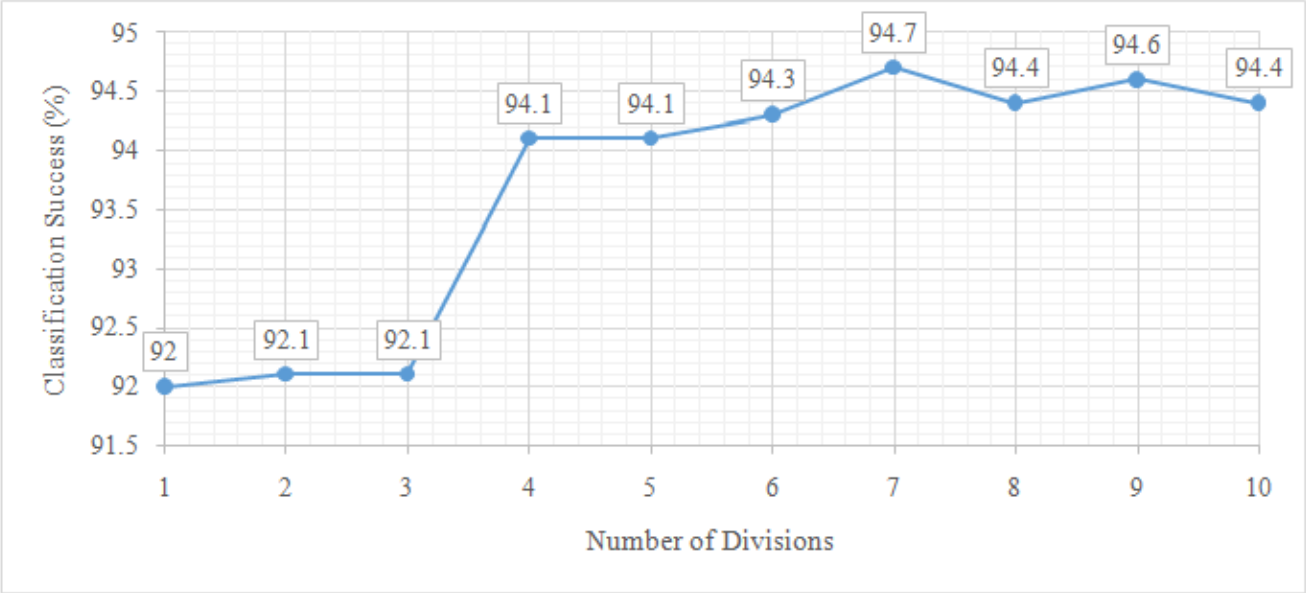

Figure 2: The Effect of Division Number on Classification Success

The error matrix of 7 segmented results classified with a success rate of 94.70% is given in

Table 5: Error Matrix of the Highest Classification Obtained with Decision Trees

		Actual Values	
		2	4
Estimated Values	2	434	7
	4	12	234

The comparison of classification performance obtained with K nearest neighbor and decision tree algorithms is given in Table 6.

Table 6: Performance Values of the Most Successful Results Obtained

criterion	KNN	Decision Tree
Truth	0.973	0.947
Sensitivity	0.974	0.945
Decisiveness	0.97	0.971
Precision	0.984	0.984
F Score	0.979	0.964

4. DISCUSSION

Being able to extract meaningful relationships in the data obtained from the methods used to diagnose breast cancer positively affects the diagnosis and treatment processes of the disease. In this study, it has been shown that these meaningful associations can be made with a high success rate with the help of k-nearest neighbor and decision tree algorithms of the breast cancer data set. The accuracy rate of 97.4% and 94.70% obtained with both algorithms show that its study will help shorten the diagnosis processes of the disease. Several studies in the literature use the breast cancer

dataset [19]. When the methods used were compared with the methods used in our study, our method's training and inference speed performed exceptionally well compared to other methods. In this respect, since it does not require high processing power, it has the potential to work very quickly with embedded computers.

In this study, the success rates of decision trees and k-nearest neighbor algorithms were compared; it has been observed that the performance tests performed with the k-nearest neighbor algorithm in the diagnosis of breast cancer give better results than the decision tree algorithm, but the decision trees work faster. With these aspects, it is predicted that machine learning techniques will provide easy-to-apply solutions to problems in the medical field.

It is possible to achieve high success in diagnosing diseases with machine learning techniques to be used in the design of decision-support systems that can support the decision-making processes of experts. The goal in machine learning methods is to classify with 100% accuracy. Our goal in future studies is to identify features with high impact and increase the classification accuracy to 100% by applying deep learning algorithms with/without tutorials to the same dataset or datasets containing different problems.

CONCLUSIONS

In general, it can be concluded that, although the exploratory analysis does not constitute a robust enough analysis to support decision-making definitively, it can be said that it is relevant in any data study since it is this analysis that delivers the first approach to the behavior of the objective attributes and variables. Note that not always a mining model can provide additional information to the exploratory analysis. In this case, it has been found that the k-means model performs a good classification of the data set in each of the databases. Still, it does not do so about the target variable, so the additional information provided by the model does not aim to improve decision-making around the malignancy of the mass. But this is not the case with the decision tree model. As stated in the results, this model provides transcendental information to the problem, beyond the results found in the exploratory analysis.It is important to note that the small number of study units can be seen reflected in models that are not very efficient or unreliable. The results obtained in the k-means model could improve if more information were available. Well, data sets made up of few variables, or even more so, with a small number of study units, can lead to finding models that are not very reliable, or it can even happen that the model cannot be implemented. In general, based on the study presented in this material, the application of the decision tree model in the characterisation of the malignancy of a breast mass is recommended over the two implemented models.

REFERENCES

1. *Harbeck, Nadia &Penault-Llorca, Frédérique& Cortes, Javier &Gnant, Michael &Houssami, Nehmat&Poortmans, Philip & Ruddy, Kathryn & Tsang, Janice & Cardoso, Fatima. (2019). Breast cancer. Nature Reviews Disease Primers. 5. 10.1038/s41572-019-0111-2.*
2. *Rajaguru, Harikumar&Chakravarthy, Sannasi. (2019). Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer. Asian Pacific journal of cancer prevention : APJCP. 20. 3777-3781. 10.31557/APJCP.2019.20.12.3777.*
3. *Medjahed, Seyyid Ahmed &Saadi, Tamazouzt&Benyettou, Abdelkader. (2013). Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. International Journal of Computer Applications. 62. 1-5. 10.5120/10041-4635.*
4. *Hadidi, moh'd&Alarabeyyat, Abdulsalam&Alhanahnah, Mohannad. (2016). Breast Cancer Detection Using K-Nearest*

Neighbor Machine Learning Algorithm. 35-39. 10.1109/DeSE.2016.8.

5. *Sarkar, Manjula& Leong, Tze-Yun. (2000). Application of K-nearest neighbors algorithm on breast cancer diagnosis problem. Proceedings / AMIA ... Annual Symposium. AMIA Symposium. 759-63.*
6. *Khorshid, Shler& Mohsin Abdulazeez, Adnan & Mohsin, Adnan. (2021). BREAST CANCER DIAGNOSIS BASED ON K-NEAREST NEIGHBORS: A REVIEW PJAEE, 18 (4)*
7. *Li, Qingbo& Li, Wenjie& Zhang, Jialin& Xu, Zhi. (2018). An improved k-nearest-neighbor method to diagnose breast cancer. The Analyst. 143. 10.1039/C8AN00189H.*
8. *Rodriguez, Victoria & Sharma, Karan & Walker, Dana. (2018). Breast Cancer Prediction with K-Nearest Neighbor Algorithm using Different Distance Measurements. 10.13140/RG.2.2.20288.79361.*
9. *Eyupoglu, Can. (2017). Breast Cancer Classification Using k-Nearest Neighbors Algorithm. , international Science and Technology Conference, July 17-19, 2017 Berlin, Germany&August 16-18, 2017 Cambridge, USA*
10. *Odajima, Katsuyoshi&Pawlovsky, Alberto. (2014). A detailed description of the use of the kNN method for breast cancer diagnosis. 688-692. 10.1109/BMEI.2014.7002861.*
11. *MurtiRawat, Ram & Panchal, Shivam& Singh, Vivek& Panchal, Yash. (2020). Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning. 534-540. 10.1109/ICESC48915.2020.9155783.*
12. *Maheshwar,. (2019). Breast Cancer Detection Using Decision Tree, Naïve Bayes, KNN and SVM Classifiers: A Comparative Study. 683-686. 10.1109/ICSSIT46314.2019.8987778.*
13. *Admassu, Tsehay&Ravulapalli, Lakshmi & Napa, Komal Kumar. (2021). Breast cancer prediction model with decision tree and adaptive boosting. IAES International Journal of Artificial Intelligence (IJ-AI). 10. 184. 10.11591/ijai.v10.i1.pp184-190.*
14. *Sathiyanarayanan, P. &Pavithra, S & Saranya, M &Makeswari, M. (2019). Identification of Breast Cancer Using The Decision Tree Algorithm. 1-6. 10.1109/ICSCAN.2019.8878757.*
15. *Osmanović, Ahmed & Abdel-Ilah, Layla &Bušatlić, Indira &Halilovic, Sabina &Tarakčija, Dževida&Mrkulic, Fatima &Hodzic, Adnan &Kevric, Jasmin. (2017). Decision Tree Classifiers for Breast Cancer Detection.*
16. *MeorBadiauzzaman, Iffah&Moey, Soo Foon&Azemin, Mohd. (2019). The use of decision tree in breast cancer related research. 8. 1344-1355. 10.35940/ijitee.I3290.0789S319.*
17. *Elsayad, Alaa&Elsalamony, Hany. (2013). Diagnosis of Breast Cancer using Decision Tree Models and SVM. International Journal of Computer Applications. 83. 19-29. 10.5120/14445-2604.*
18. *Palanisamy, Hamsagayathri&Sampath, P.. (2017). Decision Tree Classifiers For Classification Of Breast Cancer. International Journal of Current Pharmaceutical Research. 9. 31. 10.22159/ijcpr.2017v9i1.17377.*
19. *Liu, Yi & Yi, Wu. (2017). Decision Tree Model in the Diagnosis of Breast Cancer. 176-179. 10.1109/ICCTEC.2017.00046.*
20. *Rajendiran, P., et al. "Customer Relationship Management in the Manufacturing Industry, Using Data Mining Techniques." International Journal of Educational Science and Research (IJESR) 7.6 (2017): 63-70.*
21. *Singh, Gurpreet, and Er Rajwinderkaur. "Review Paper On Decision Tree Data Mining Algorithms To Improve Accuracy In Identifying Classified Instances Using Large Dataset." International Journal Of Computer Science Engineering And Information Technology Research (Ijcseitr) 7 (2017): 67-70.*
22. *Hiremath, Basavaraj, and S. C. Prasannakumar. "Automated evaluation of breast cancer detection using SVM classifier." International Journal of Computer Science Engineering 5.1 (2015): 7-16.*

23. *Moffatt, Stanley, and Mutune Wangari. "Enhanced Silencing of Bmi-1 and Htert Gene Expression with Ngr-Pei-Coupled Sirna Lipoprotein Nano-Complexes." TJPRC: Journal of Medicine and Pharmaceutical Science (TJPRC: JMPS) 2 (2016): 9-18.*